کنترل ترافیک یک چهارراه راهنمایی رانندگی با استفاده از الگوریتم‌های یادگیری تقویتی (یادگیری-Q، سارسا و مسیرهای شایستگی)

علیرضا عربی*، کارشناسی ارشد، دانشکده برق و مهندسی پزشکی، دانشگاه صنعتی سجاد، مشهد، ایران
امین نوری، مربی، دانشکده برق و مهندسی پزشکی، دانشگاه صنعتی سجاد، مشهد، ایران
*پست الکترونیکی نویسنده مسئول alireza.arabi1@sadjad.ac.ir
دریافت: ۹۶/۰۹/۰۸- پذیرش: ۹۷/۰۱/۱۸
صفحه۵۵-۶۸

چکیده

یکی از مهمترین اهداف پژوهش در حوزه حمل و نقل، بهینه کردن جریان‌های ترافیک است. امروزه با افزایش وسایل نقلیه به طور پیوسته، محدودیت در منابع ارائه شده توسط زیر ساخت‌های فعلی و ماهیت غیرخطی، پویا و تصادفی بودن جریان ترافیک، استفاده از روش‌های هوشمند در کنترل ترافیک به خصوص روش‌های حل مساله یادگیری تقویتی حائز اهمیت است. روش یادگیری تقویتی علاوه بر سادگی و نداشتن پیچیدگی محاسباتی، در عمل بی نیاز به مدل ریاضی محیط می‌باشد و خاصیت تطبیق پذیری با شرایط محیط و مقاوم بودن در برابر تغییرات محیطی را دارد. در این مقاله کنترل ترافیک یک تقاطع با سه روش از زیر روش‌های حل مساله یادگیری تقویتی (یادگیری-Q، سارسا و مسیرهای شایستگی) انجام شده است. نتایج شبیه‌سازی حاکی از آن است که مسیرهای شایستگی یک روش کنترلی بروزتر و بهینه‌تر نسبت به دو روش یادگیری-Q و سارسا که پیشتر در مقالات کنترل ترافیک مورد استفاده قرار گرفته است، می‌باشد.

واژه‌های کلیدی: کنترل سیگنال ترافیک، یادگیری تقویتی، یادگیری-Q، ساراسا، مسیرهای شایستگی

۱- مقدمه

چراغ‌های راهنمایی جزء آشناترین ابزارها در سیستم کنترل ترافیک به حساب می‌آید. این ابزار در تقاطع‌های موجب تنظیم عبور و مرور و افزایش ایمنی وسایل نقلیه می‌شود. در (Liu ,2007) اولین بار برای کنترل یک تقاطع مستقل از کنترل کننده منطق فازی استفاده شده است. در این کنترلر فازی، از مدل ماکروسکوپیک برای شبیه‌سازی طول صف و محاسبه تأخیر استفاده شده است. نتایج حاکی از آن است، این سیستم کنترلی نسبت به ثابت زمانی عملکرد رضایت‌بخشی دارد.

در (Qiao, Yang and Gao, 2011) یک کنترل‌کننده منطق فازی دو قسمتی برای کنترل جریان ترافیکی تقاطع‌ها استفاده شده است. نحوه کار این سیستم به این صورت است که در قسمت اول، دوزمانه یا چهارزمانه بودن چراغ راهنمایی

مشخص می‌شود، در قسمت دوم مدت زمان فاز سبز مشخص می‌شود. در مرجع (Park, Messeand and Urbanik, 2000) از الگوریتم ژنتیک برای هوشمندسازی عملکرد چراغ راهنمایی استفاده شده است. این الگوریتم توانایی بهینه‌سازی فاز، طول سیکل، آفست‌ها و تقاضای عبوری به صورت همزمان را داراست. امروزه روش‌های مبتنی بر الگوریتم ژنتیک در سیستم‌های کنترل ترافیک شهری استفاده می‌شود.

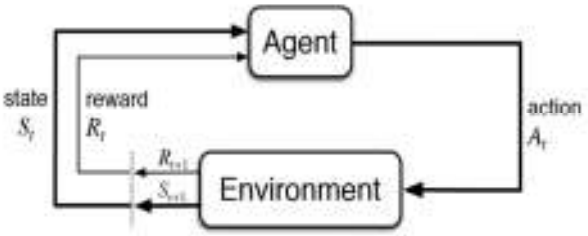
در (Abhishek and Bihari Misra 2016) یک مدل ترکیبی متشکل از الگوریتم ژنتیک و تاخیر زمانی شبکه‌های عصبی جهت پیش‌بینی جریان ترافیک در کوتاه مدت مورد بررسی قرار گرفته است. همچنین الگوریتم ژنتیک بهبود یافته جهت برنامه‌ریزی هوشمند ناوگان اتوبوسرانی در سطح شهر در (Chang-sheng, Jian-bo and Wen-yi 2010) مورد استفاده قرار گرفته است. نتایج نشان می‌دهد که الگوریتم ژنتیک بهبود یافته می‌تواند بهترین نتیجه تقریبی در فضای جستجو عظیمی از بهینه‌سازی را پیدا کند، در حالی که تا حد زیادی بهره‌وری ناوگان حاصل شود. در (Tianshu Chu, Shuhui Qu and Jie Wang 2016) روشی جهت کنترل چراغ‌های راهنمایی مبتنی بر سیستم‌های چند عامله ارائه شده است. در این روش، عامل‌ها از الگوریتم یادگیری-Q، جهت کنترل سیگنال ترافیک برای شبکه‌های ترافیک در مقیاس بزرگ استفاده می‌کنند. روش غیر متمرکز چند عاملی مبتنی بر یادگیری تقویتی در (Dusparic, Monteil and Cahill 2016) معرفی شده است تا خاصیت مقیاس‌پذیری و تطبیق‌پذیری زمان حقیقی را در سیستم‌های کنترل ترافیک شهری افزایش دهند.

در (Prabuchandran and Kumar 2016) فرایند تصمیم‌گیری مارکوف و اعمال الگوریتم یادگیری تقویتی جهت یافتن خط مشی بهینه استفاده شده است.

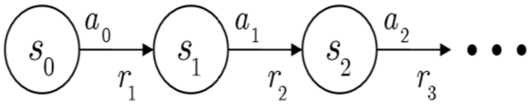
در (Lurong, Zhang and Shi 2010) از یک کنترلر فازی در کنترل هماهنگ سیگنال ترافیک برای چند تقاطع هم‌جوار استفاده شده است. کنترلر فازی با دریافت اطلاعات از یک سه راه راهنمایی و رانندگی طراحی شده است. کنترلر مذکور در ۱۲ فاز از به وجود آمدن شرایط ترافیک سنگین جلوگیری می‌نماید. در (Kim, 1994) سعی شده است، یک شبکه متشکل از چند تقاطع کنترل شود، در نتیجه کنترل‌کننده طراحی شده حداکثر می‌تواند تا چهار فاز را پشتیبانی کند. تقاطع‌های مورد بررسی، دو مسیر حرکت مستقیم و یک مسیر برای حرکت گردش به چپ لحاظ شده است. در این سیستم کنترلی برای هر فاز ابتدا یک مدت زمان سبز ثابت (هفت ثانیه) در نظر گرفته شده است. این مدت زمان تا ۵۷ ثانیه قابلیت تمدید دارد. در (Odeh, S, M 2015) ترکیبی از الگوریتم ژنتیک و فازی جهت بدست آوردن عملکرد بالاتر نسبت به کنترلر مبتنی بر منطق فازی کلاسیک استفاده شده است. نتایج شبیه‌سازی ۳۱ درصد بهبود عملکرد سیستم کنترلی را نسبت به کنترلر منطق فازی معمولی را نشان می‌دهد. در (Tian, Y., Li, Z., Zhou, D., Song, J., & Xiao, D, 2008) از ترکیب شبکه عصبی و منطق فازی استفاده شده است. در این ترکیب علاوه بر افزایش دقت عملکرد کنترل‌کننده، بهینه‌سازی قواعد فازی نیز ایجاد شده است. در (Xu Dongling, et al, 1992) از یک کنترلر فازی جهت هماهنگ‌سازی چراغ‌های راهنمایی برای یک شبکه متشکل از چند تقاطع استفاده شده است. کنترلر فازی با دریافت اطلاعات از تقاطع‌ها مانع مسدود شدن و ایجاد وضعیت بحرانی می‌شود.

۲- یادگیری تقویتی

مفهوم یادگیری، به سه دسته یادگیری نظارت شده، یادگیری غیر نظارت شده و یادگیری تقویتی تقسیم می‌شود. در یادگیری نظارت شده، ناظر یا مربی وجود دارد که با دانستن بهترین عمل در هر حالت یا وضعیت، برای بهبود عملکرد عامل یادگیرنده توصیه‌هایی را مطرح می‌کند. در یادگیری غیر نظارت شده که نقطه روبه‌روی یادگیری نظارت شده، می‌باشد، هیچ معلم یا ناظری وجود ندارد و عامل یادگیرنده، باید براساس آنچه دریافت کرده به آنچه پیشتر دیده است به نوعی ربط دهد. اما در یادگیری تقویتی، عمل یادگیری برپایه بازخوردهایی است که خود براساس عبارت‌های کمکی مثبت یا منفی می‌باشد. در این نوع یادگیری معلم به طور مستقیم بهترین عمل در هر وضعیت را به عامل یادگیرنده نمی‌گوید و فقط به عامل، میزان خوب بودن یا بد بودن عمل گفته می‌شود.



شکل ۱. شمای کلی یادگیری تقویتی[Sutton and Barto 1998]



شکل ۲. مجموعه‌ای از حالت، عمل و پاداش‌های متوالی

در یک مسئله یادگیری تقویتی با عاملی رو به رو هستیم که از طریق سعی و خطا با محیط تعامل کرده و یاد می‌گیرد تا عملی بهینه را برای رسیدن به هدف انتخاب کند.

همانطور که در شکل ۱ نشان داده شده است، عامل به طور مستقل با محیط تعامل کرده، یاد می‌گیرد و کسب تجربه می‌کند و پاداشی را از محیط دریافت می‌کند. در یادگیری تقویتی عامل‌ها دارای حسگرهایی هستند که می‌توانند اطلاعاتی جهت تعیین وضعیت خود از محیط را دریافت کنند. به طور کلی، تمام عامل‌های یادگیری تقویتی مکانیزم ثابتی دارند، می‌توانند محیط خود را بدون داشتن مدل ریاضی درک کنند، اعمالی را انتخاب کنند و محیطشان را تحت تاثیر قرار دهند. در مساله یادگیری تقویتی به دنبال بیشینه کردن تابع ارزش برای تمام حالت‌های تعریف شده‌ی سیستم می‌باشیم، بنابراین طبق شکل ۲ وقتی از حالت s شروع می‌کنیم تحت سیاست π گام برمی‌داریم آنگاه به دنبال آن هستیم با انجام کنش a ارزش دریافتی تحت انجام یک سری اعمال را بیشینه کنیم و وارد حالت جدید شویم و ارزش یک حالت بر اساس سیاست π به صورت زیر تعریف می‌شود:

$$V^{\pi}(\mathrm{s}) = \mathrm{E}_{\pi}\{R_t|S_t = s\} = \mathrm{E}_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\,|S_t = s\right\} \quad (۱)$$

در معادله (۱) $\mathrm{R_t}$ مجموع پاداش‌های دریافتی با شروع از حالت **s** می‌باشد و γ نرخ فراموشی می‌باشد. یکی از خواص بنیادی توابع ارزش به دست آوردن رابطه (۲) به صورت بازگشتی، می‌باشد که تحت عنوان معادله بلمن، شناخته می‌شود.

$$v^{\pi}(s) = \sum_a \pi(s,a)\sum_{s'} p^a_{ss'}\left[R^a_{ss'} + \gamma v^{\pi}(s')\right] \quad (۲)$$

در معادله (۲)، s حالت فعلی a عمل فعلی s' حالت بعدی می‌باشد. رابطه فوق نشان دهنده رابطه میان ارزش حالت فعلی و ارزش حالت بعدی است.

۲-۱- روش‌های یادگیری تقویتی

روش‌های یادگیری تقویتی به سه دسته برنامه‌ریزی پویا، مونت کارلو و تفاوت‌گذرا تقسیم می‌شود. در ادامه به بررسی سه روش تفاوت گذرا (مسیرهای شایستگی، الگوریتم Q و الگوریتم سارسا) می‌پردازیم.

۲-۱-۱ لگوریتم سارسا

این الگوریتم در سال ۱۹۹٤ مطرح گردید و یکی از روش‌های پایه تفاوت گذرا می‌باشد که بر پایه سیاست رفتاری عمل می‌کند. بدین معنا که مسیری که گام بر می‌دارد با مسیری که بروزرسانی می‌کند، یکسان می‌باشد. عامل در حالت جاری

یک عمل را انتخاب می‌کند و به حالت جدید می‌رود و پاداش دریافت می‌کند و سپس عمل بعدی را انتخاب می‌کند . شکل ۳ نشان دهنده ساختار کلی به روز رسانی و نحوه عملکرد این الگوریتم می‌باشد.

1. Initialize Q(s,a) arbitarily
2. Repeat (for each episode)
3. Initialize s
4. Choose a from s using policy derived from Q (e.g, ε-greedy)
5. Repeat (for each step of episode)
6. Take a action, observe r, s'
7. Choose a' from s' using policy derived from Q (e.g, ε-greedy)
8. $Q(s_t,a_t) = Q(s_t,a_t) + \alpha(r_{t+1} + Q(s_{t+1},a_{t+1}) - Q(s_t,a_t))$
9. $s \rightarrow s'$, $a \rightarrow a'$
10. Until s is terminal

شکل ۳. شبه کد سارسا

1. Initialize *Q(s,a)* arbitarily
2. Repeat (for each episode)
3. Initialize *s*
4. Repeat (for each step of episode)
5. Choose a from s using policy derived from *Q (e.g, ε-greedy)*
6. Take action a, observe *r, s'*
7. $Q(s_{t+1},a_{t+1}) = Q(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_{a'} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t))$
8. $s \leftarrow s'$;
9. Until *s* is terminal;

شکل ٤. شبه کد یادگیری-Q

طبق شکل ۳ ابتدا مقادیر Q را برای همه حالت- عمل‌ها برابر صفر قرار می‌دهیم، سپس در هر اپیزود با استفاده از یک سیاست نرم بر پایه Qهای تخمین زده شده تا آن لحظه یک عمل را انتخاب می‌کنیم و انجام می‌دهیم. پاداش لحظه‌ای دریافت می‌شود و به یک حالت جدید می‌رویم. در این حالت جدید نیز عمل دلخواه را انتخاب می‌کنیم و ارزش حالت-عمل اولیه را بر اساس پاداش لحظه‌ای آن و ارزش حالت-عمل بعدی طبق معادله (۳) بروز رسانی می‌کنیم[۲٤].

$$Q(s_{t+1},a_{t+1}) = Q(s_t,a_t) + \alpha[r_{t+1} + \gamma * Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \quad (۳)$$

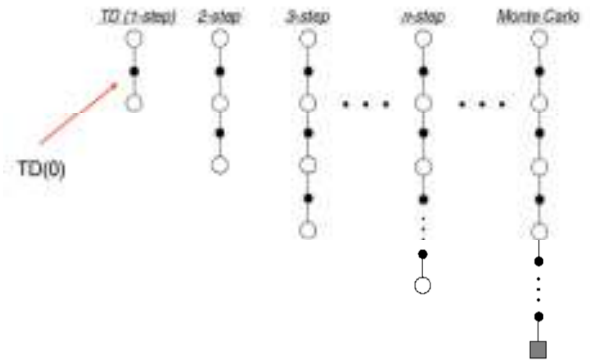
۲-۱-۲ الگوریتم یادگیری Q

الگوریتم یادگیری Q که توسط واتکینز در سال ۱۹۹۸ برای حل مسائل تصمیم‌گیری مارکوف با اطلاعات ناقص معرفی شد، در دسته روش‌های تفاوت گذرا قرار می‌گیرد. این الگوریتم نیاز به مدل دقیق و شفاف محیط ندارد و می‌تواند از طریق تجربه‌ای که در اثر تعامل با محیط بدست می‌آورد، برای یافتن استراتژی بهینه استفاده شود. پاداش عددی r_n که به یک زوج حالت و عمل تعلق دارد. بنابراین مقدار این پاداش تابعی از حالاتی است که سیستم در آن قرار دارد و عملی است که عامل در این حالت انجام می‌دهد. در یادگیری Q تابع مقدار با استفاده از یک جدول حالت و عمل مشخص می‌شود که مقدارها به صورت Q تعریف می‌شود فرض کنید $x=\{x_1,x_2,\ldots,x_n\}$ مجموعه از k حالت قابل قبول محیط و $A=\{a_1,a_2,\ldots,a_m\}$ مجموعه ای از m عمل قابل قبول باشد که توسط عامل قابل انجام باشد که در آن قرار دارد درک کند. در رابطه ۳ تنها مقدارQ متعلق به زوج حالت و عمل است. ضریب α نرخ یادگیری است که در باز بین صفر و یک قرار دارد و مقادر کم آن به منظور بهره-برداری کم از اطلاعات جدید به دست آمده است. پاداش دریافتی در حالت بعد با نرخ کاهشی γ در زمان حال محاسبه می‌شود. جهت در نظر گرفتن ارزش زیاد برای پاداش مورد نیاز در آینده باید مقدار ضریب را افزایش داد.

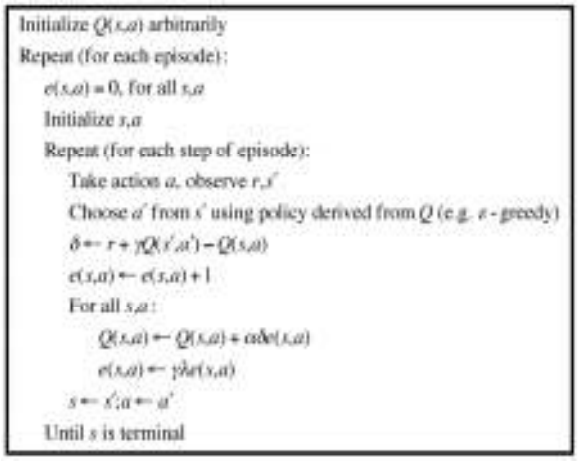
طبق شکل ٤ ابتدا مقادیر Q را برای همه حالت عمل‌ها برابر صفر قرار می‌دهیم. سپس در هر اپیزود با استفاده از یک سیاست نرم بر پایه Qهای تخمین زده شده تا آن لحظه یک عمل را انتخاب می‌کنیم و انجام می‌دهیم. پاداش لحظه‌ای دریافت می‌شود و به یک حالت جدید می‌رویم. باید توجه داشت که یادگیری-Q یک روش off-policy است، یعنی عامل با یک سیاست نرم در محیط در حال زندگی است در حالیکه همزمان یک سیاست حریصانه دیگر را یاد می‌گیرد و بهینه می‌کند. در این روش برخلاف سارسا ارزش حالت عمل مشاهده شده را بر اساس ارزش حالت عمل بعدی بروز نمی‌کنیم بلکه بروزرسانی بر اساس ارزش عملی که در حالت بعد بیشترین ارزش را داراست صورت می‌گیرد [Sutton and Barto 1998].

۲-۱-۳ مسیرهای شایستگی

مسیرهای شایستگی یکی از مکانیزم‌های اصلی تقویت یادگیری است. در مسیرهای شایستگی دو حالت مشاهده اثر وجود دارد. از نظر تئوری، مسیرهای شایستگی پلی بین روش مونت کارلو و تفاوت گذرا است (دیده به جلو) و از دیدگاه دیگر دارای دید به عقب است، در واقع مسیرهای شایستگی دارای حافظه است و می‌تواند کنش‌ها و حالت‌ها را ذخیره کند [Sutton and Barto 1998].



شکل ۵. حالت های مختلف دید به جلو[Sutton and Barto 1998]


Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode):
$e(s,a) = 0$, for all s,a
Initialize s,a
Repeat (for each step of episode):
Take action a, observe r,s'
Choose a' from s' using policy derived from Q (e.g. ε - greedy)
$\delta \leftarrow r + \gamma Q(s',a') - Q(s,a)$
$e(s,a) \leftarrow e(s,a) + 1$
For all s,a:
$Q(s,a) \leftarrow Q(s,a) + \alpha\delta e(s,a)$
$e(s,a) \leftarrow \gamma\lambda e(s,a)$
$s \leftarrow s'; a \leftarrow a'$
Until s is terminal

شکل ٦. شبه کد مسیرهای شایستگی سارسا[Sutton and Barto 1998]

در شکل ٥ حالت‌های مختلف دید به جلو در روش تفاوت گذرا، مسیرهای شایستگی و مونت کارلو نشان داده شده است. در روش تفاوت گذرا عامل در هر گام (حالت فعلی) با انجام کنش و مشاهده حالت بعدی می‌تواند خود را بروز کند اما در مونت کارلو عامل زمانی می‌تواند خود را بروز کند که به هدف برسد. در مسیرهای شایستگی عامل می‌تواند تعداد گام‌های بروزرسانی را خود انتخاب کند.

در شکل ٦ ابتدا به تعداد حالت‌ها و کنش‌ها، ماتریس ارزش حالت-عمل و ماتریس خطای حالت-عمل را تشکیل می‌دهیم سپس حالت فعلی را مشاهده می‌کنیم و بر اساس سیاست حریصانه عملی را انتخاب می‌کنیم. با انجام کنش انتخابی و مشاهده حالا بعدی و دریافت پاداش، کنش حالت جدید را بر اساس سیاست فوق انتخاب می‌کنیم سپس مجموع پاداش دریافتی و ارزش حالت عمل جدید را ارزش حالت-عمل فعلی کم می‌کنیم. بر اساس نگاه به عقب مسیرهای شایستگی اگر حالت فعلی که در آن هستیم قبلا مشاهده کرده باشیم، به میزان خطای حالت عمل یک واحد می‌افزاییم، سپس بر اساس ارزش حالت عمل فعلی، نرخ یادگیری و خطای حالت عملی فعلی ارزش حالت-عمل فعلی را بروز می‌کنیم و خطای حالت عمل فعلی را بر اساس نرخ فراموشی و نرخ تعادلی بین مونت کارلو و تفاوت گذرا بروز می‌کنیم. در نهایت حالت و عمل جدید را به عنوان حالت و عمل فعلی در نظر می‌گیریم. در ادامه می-خواهیم از یادگیری تقویتی برای هوشمند سازی عملکرد چراغ راهنمایی استفاده کنیم[Sutton and Barto 1998].

۳- مدل‌سازی

برای بررسی نتایج عملکرد این سه روش، سه مدل چهارراه شبیه‌سازی شده است که هر محیط دارای عملکردی با یکی از روش‌های یادگیری تقویتی می‌باشد. از ویژگی‌های این مدل می‌توان به تصادفی و پویا بودن و متغییر با زمان بدون آن اشاره کرد. در واقع میزان خروجی خودروها از هر خیابان بسته به اهمیت آن خیابان کاملا تصادفی هست و همچنین ورودی‌های خودرو به هر خیابان بسته به اهمیت آن خیابان کاملا تصادفی و متفاوت نسبت به هم است. برای اطمینان از صحیح بودن نتایج در هر سه محیط تمامی پارامترها اعم از خروجی و ورودی خودرو به خیابان‌ها باهم برابر است. تفاوتی این سه محیط با یکدیگر دارند در انتخاب اعمال است، در واقع در محیط یک انتخاب عمل بر اساس الگوریتم مسیرهای شایستگی، در محیط دوم انتخاب عمل بر اساس الگوریتم Q و در محیط سوم انتخاب عمل بر اساس الگوریتم سارسا است.

۳-۱ تعریف حالات و اعمال و پاداش برای محیط ترافیک

۱-۱-۳- حالت‌ها: هر خیابان به ۵ قسمت تقسیم کردیم که در قسمت می‌تواند ۳ خودرو جا دهد. علت این امر کاهش تعداد حالات سیستم می‌باشد. در کل ٦٢٥ حالت مورد بررسی قرار می‌گیرد.

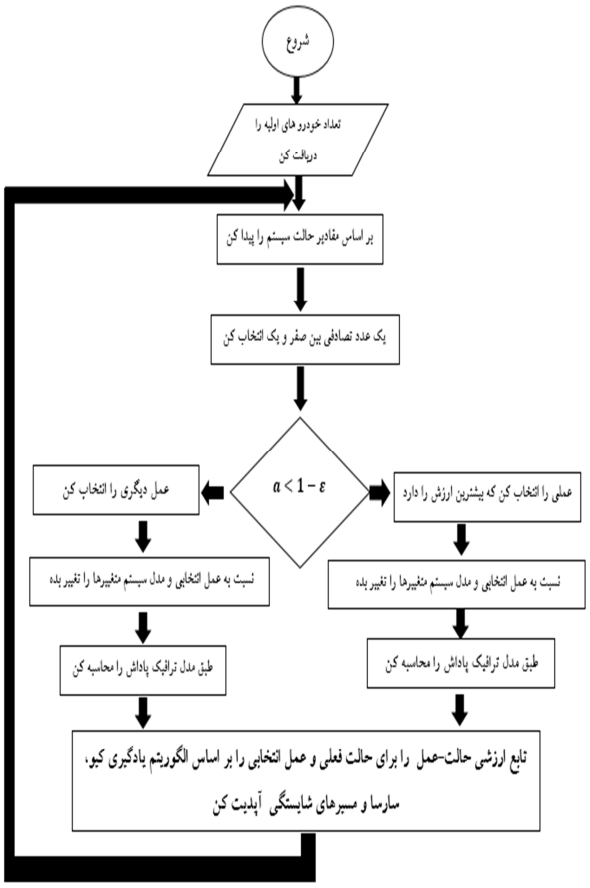
۲-۱-۳- اعمال: در این پژوهش ٢٤ عمل برای عامل یادگیرنده در نظر گرفته شده است. هر عمل دارای ۲ بعد زمانی و موقعیتی است، در بعد زمانی دارای ٤ حالت، ۳۰ ثانیه،٦٠ ثانیه، ۹۰ ثانیه و ۱۲۰ ثانیه می‌باشد و در بعد موقعیتی دارای ٦ حالت، سبز شدن چراغ‌راهنمایی خیابان شمالی، سبز شدن چراغ‌راهنمایی خیابان شرقی، سبز شدن چراغ‌راهنمایی خیابان جنوبی، سبز شدن چراغ‌راهنمایی خیابان غربی، سبز شدن همزمان چراغ‌راهنمایی خیابان شمالی و جنوبی و سبز شدن همزمان چراغ‌راهنمایی خیابان شرقی و غربی، می‌باشد.

۳-۱-۳- تابع پاداش: بر اساس تعداد خودروها و زمان تاخیر در تقاطع تعریف شده است. بدین ترتیب که هر چقدر تعداد خودروهای متوقف پشت چراغ قرمز و زمان تاخیر بیشتر شود مقدار تابع هزینه بیشتر می‌شود. تابع هزینه بصورت معادله (٤) بیان شده است که ترکیبی از اندازه صف هر یک از چهار خیابان و زمان انتظار تجمعی از همه اتومبیل‌های متوقف شده پشت چراغ قرمز می‌باشد.

$$\sum_{i=1}^{4} \beta_q (q_i)^{\theta_q} + \beta_\omega (\omega_i)^{\theta_\omega} \qquad (٤)$$

که در آن q_i اندازه صف هر یک از چهار خیابان و w_i زمان انتظار تجمعی از همه اتومبیل‌های متوقف است. که این دو شاخص عملکرد تا حد زیادی همبسته می‌باشد. می‌خواهیم یک موازنه بین این دو مورد برقرار کنیم، بگونه‌ای که یک خودرو به مدت زمان بینهایت در صف نماند. بر این اساس با توجه به آزمایشات و مشاهدات $\beta_q = 1$ و $\beta_\omega = 2$ در نظر گرفته شده‌اند تا q_i و w_i تقریبا در یک مقیاس قرار گیرند و θ_q و θ_w برابر ۱/۵ در نظر گرفته شده‌اند. در این مقاله برای انتخاب عمل از روش $\varepsilon - greedy$ استفاده شده است. معادله بروزرسانی الگوریتم یادگیری Q طبق معادله (۵) است.

معادله بروزرسانی الگوریتم یادگیری سارسا طبق معادله (٦) است و معادله بروزرسانی الگوریتم یادگیری مسیرهای شایستگی طبق معادله (۷) است. که در این روابط Q، تابع ارزش بر اساس حالت (تعدادخودروهای چهارراه) و عمل (نحوه انتخاب چراغ سبز و زمان روشن بودن)، ns، شماره حالت فعلی سیستم (بر اساس تعداد خودروها در چهارراه)، sa، شماره عمل انتخابی فعلی (شماره چراغ رهنمایی و زمان سبز اختصاص داده شده)، nns، شماره حالت جدید سیستم (بر اساس تعداد خودروها در چهارراه بعد از اعمال تاثیر کنش انتخابی و ورود خودروهای) و nsa، شماره عمل انتخابی در حالت جدید است. Qval، ارزش حالت–عملی فعلی است. new_Qval، ارزش حالت عمل بعدی است و e_trace خطای حالت–عمل فعلی است.



شکل ۷. فلوچارت سه الگوریتم یادگیری تقویتی در مدل ترافیکی

٤-نتایج

$$Q(ns, sa) = Q(ns, sa) + 0.5[\sum_{i=1}^{4} \beta_q (q_i)^{\theta_q} + \beta_\omega (\omega_i)^{\theta_\omega} + 0.5 * \max Q(nns, nsa) - Q(ns, sa)] \quad (٥)$$

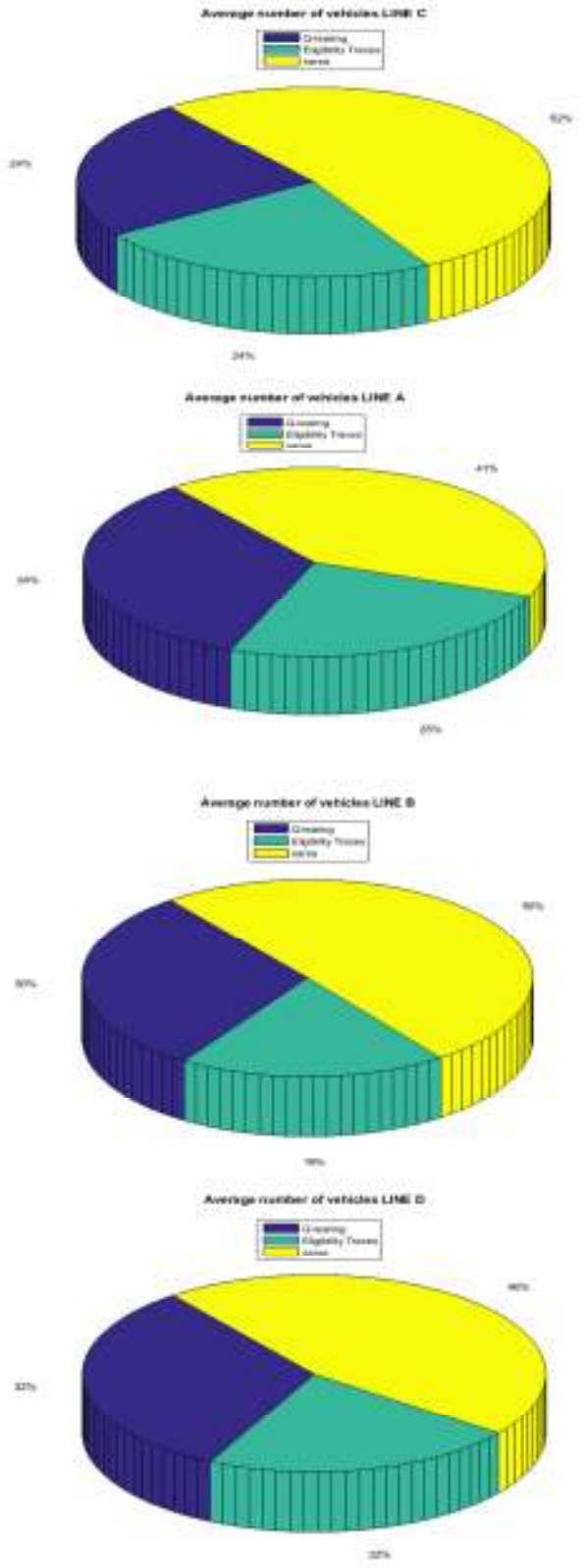
$$Q(ns, sa) = Q(ns, sa) + 0.5[\sum_{i=1}^{4} \beta_q (q_i)^{\theta_q} + \beta_\omega (\omega_i)^{\theta_\omega} + 0.5\, Q(nns, nsa) - Q(ns, sa)] \quad (٦)$$

$$Q(ns,sa) = Q(ns,sa) + 0.5 * delta * e\_trace(ns,sa) \quad (٧)$$

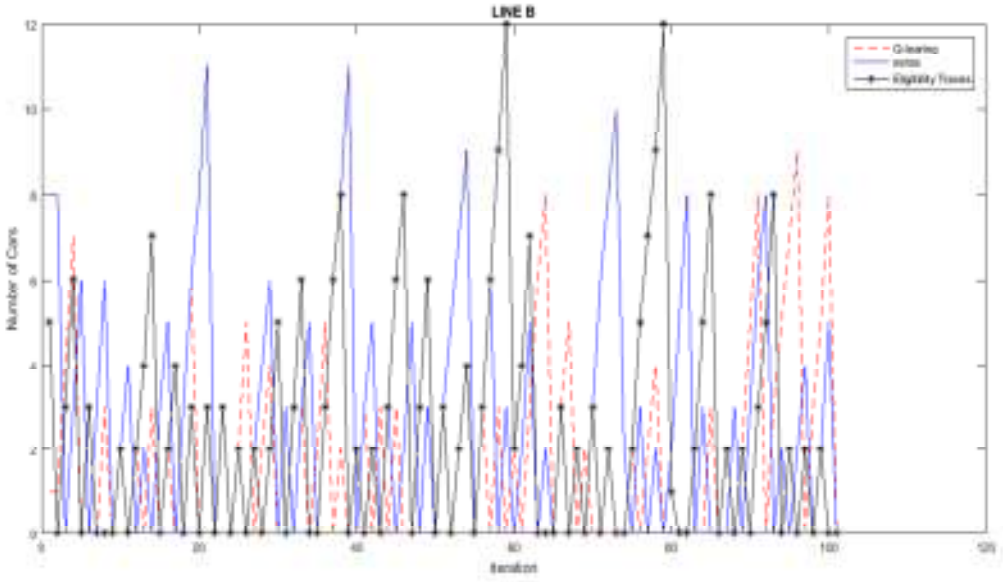
$$delta = (\sum_{i=1}^{4} \beta_q (q_i)^{\theta_q} + \beta_\omega (\omega_i)^{\theta_\omega} + 0.9 * new\_Qval - Qval) \quad (٨)$$

در کنترل ترافیک چندین جریان ترافیکی برسر زمان و فضا رقابت می‌کنند اما اغلب معیارهای متفاوتی برای جریان‌های ترافیکی وجود دارد. معمولاً معیار زمان تاخیر پارامتر تعیین کننده در کارایی سیستم‌های کنترل ترافیک به شمار می‌رود.

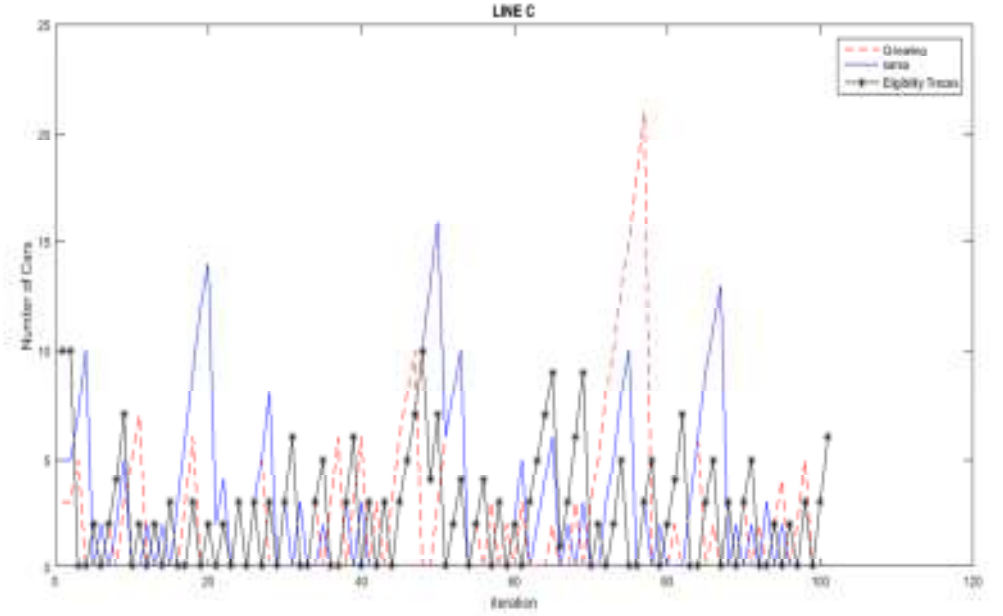
شکل ۸ میزان متوسط تعداد خودروها در چهار خیابان شمالی جنوبی شرقی و غربی است که قسمت آبی رنگ مربوط به سیستم با الگوریتم مسیرهای شایستگی است و قسمت سبز رنگ مربوط سیستم با الگوریتم یادگیری-Q است و قسمت زرد رنگ مربوط به سیستم الگوریتم یادگیری سارسا است.



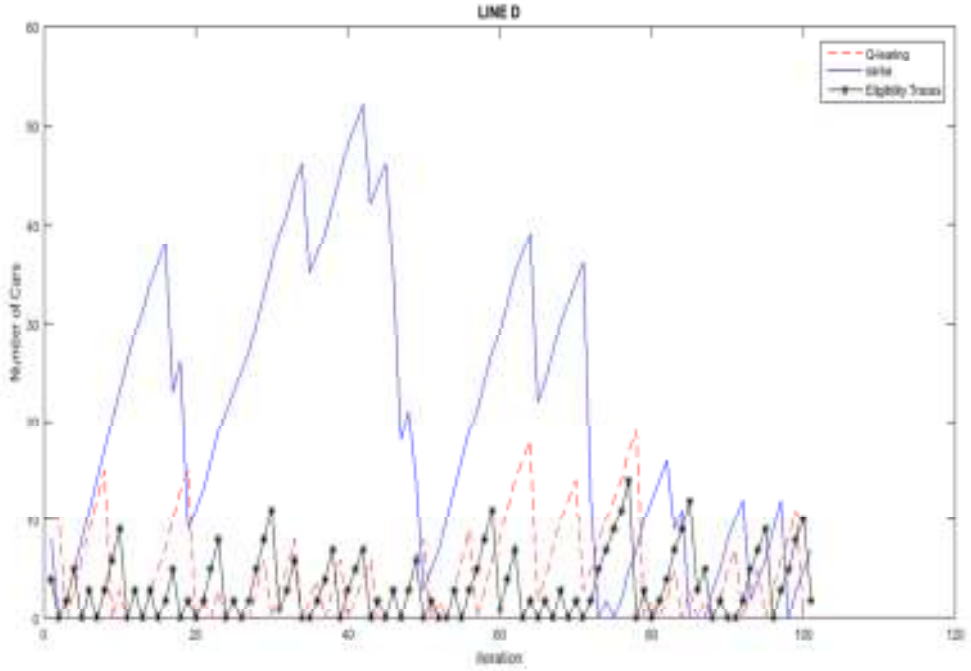
شکل ۸. مقایسه میزان متوسط تعداد خودرو چهار خیابان با سه الگوریتم یادگیری تقویتی



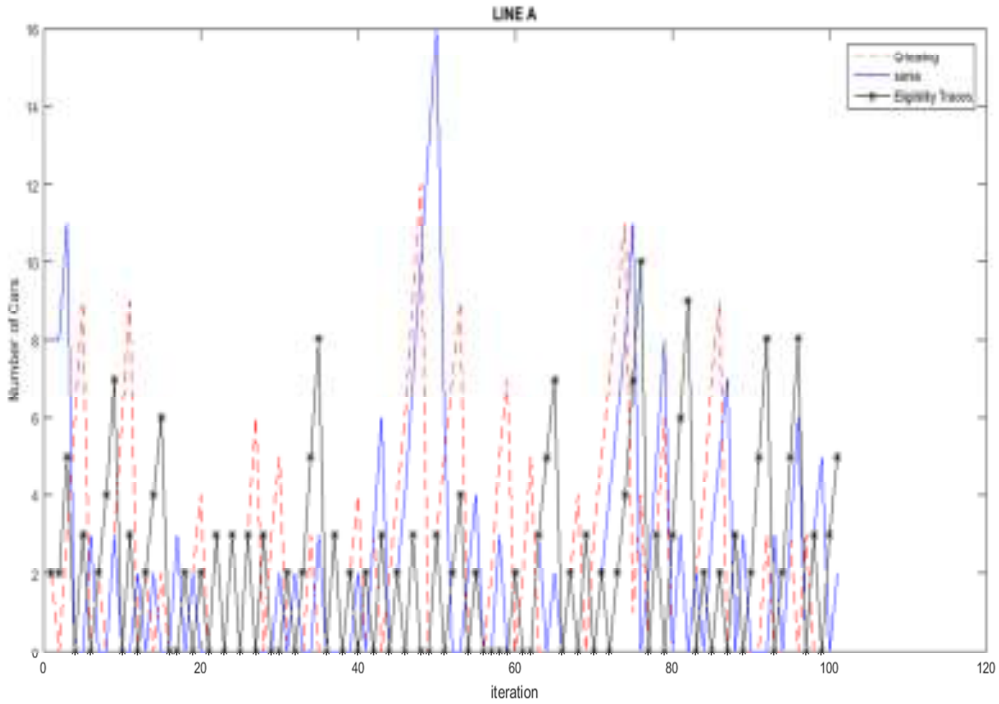
شکل ۹. مقایسه نحوه عملکرد سه الگوریتم در عمل‌های انتخابی و تعداد خودروها در خیابان شرقی



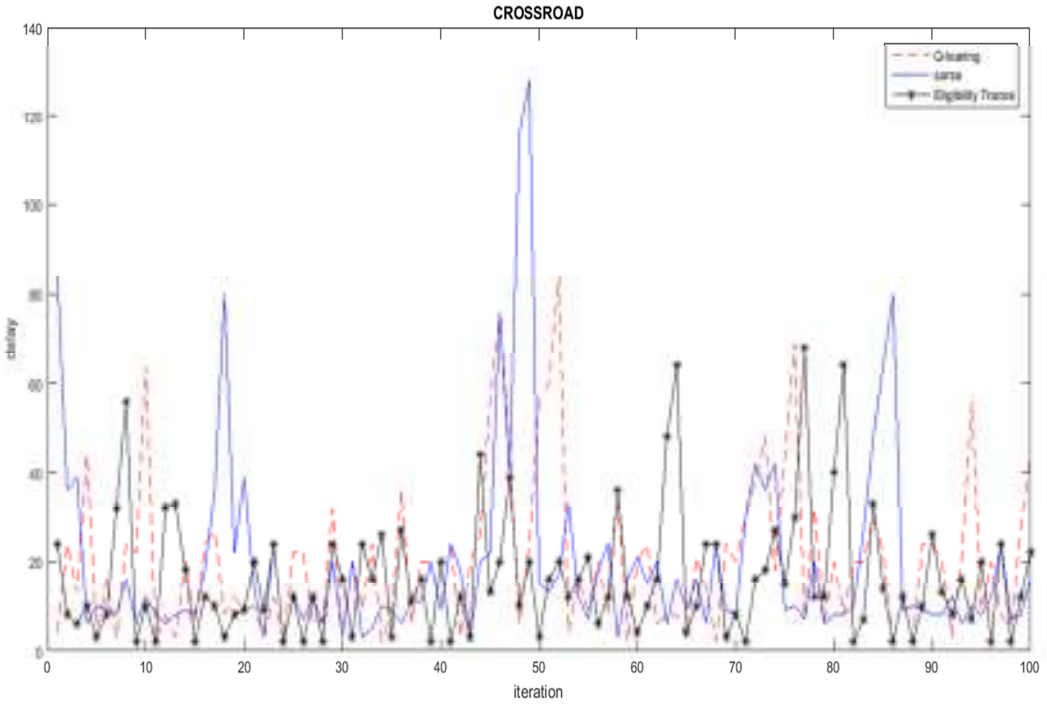
شکل ۱۰. مقایسه نحوه عملکرد سه الگوریتم در عمل‌های انتخابی و تعداد خودروها در خیابان جنوبی



شکل ۱۱. مقایسه نحوه عملکرد سه الگوریتم در عمل‌های انتخابی و تعداد خودروها در خیابان غربی



شکل ۱۲. مقایسه نحوه عملکرد سه الگوریتم در عمل‌های انتخابی و تعداد خودروها در خیابان شمالی

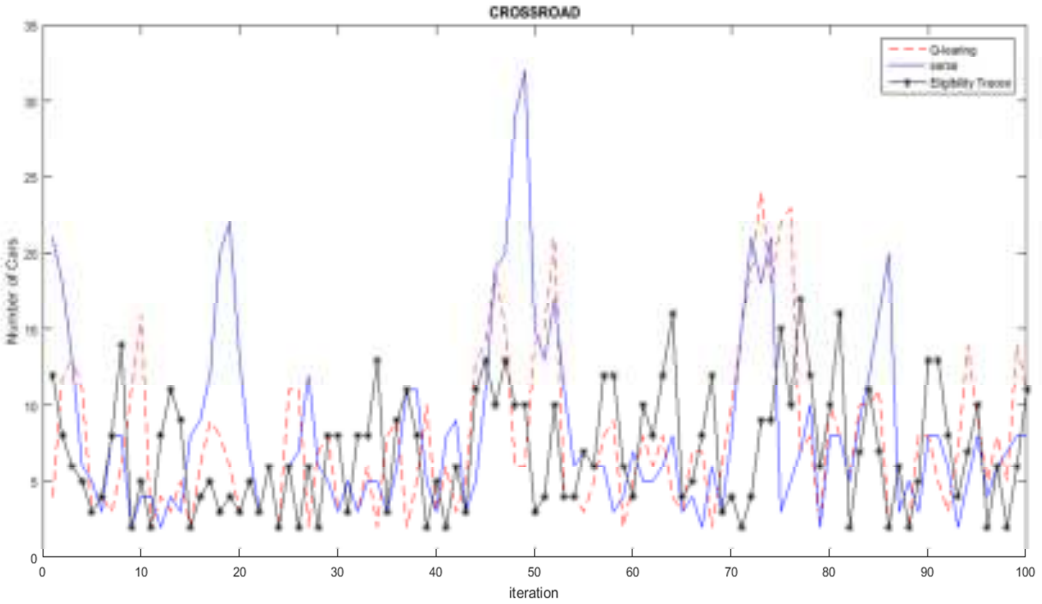


شکل ۱۳. مقایسه سه الگوریتم در میزان تاخیر خودروها پشت چراغ قرمز کل چهارراه

شکل‌های ۹، ۱۰، ۱۱ و ۱۲ میزان تعداد خودروها در هر گام زمانی در چهار خیابان شمالی، جنوبی، شرقی و غربی است که نمودار آبی رنگ مربوط به سیستم با الگوریتم سارسا است. نمودار قرمز رنگ مربوط سیستم با الگوریتم یادگیری Q است و نمودار سیاه رنگ مربوط به سیستم با الگوریتم مسیرهای شایستگی است. متوجه می‌شویم الگوریتم یادگیری مسیرهای شایستگی عملکرد بهتری دارد.

در شکل ۱۳ نمودار آبی رنگ مربوط به سیستم با الگوریتم سارسا است، نمودار سیاه رنگ مربوط به سیستم با الگوریتم مسیرهای شایستگی است و نمودار قرمز رنگ مربوط به سیستم با الگوریتم-Q است. این سیستم در مدت 100 گام مورد آزمایش قرار گرفته است که سیستم در ابتدا درحالت جستجو بوده و تعداد خودروها پشت چراغ افزایش می‌یابد اما با افزایش یادگیری میزان تعداد خودروها به مراتب کاهش می‌یابد.

طبق شکل ١٤ متوجه می‌شویم کنترل ترافیک با یادگیری تقویتی از نوع مسیرهای شایستگی عملکرد بهتری در کاهش میزان تاخیر خودروها پشت چراغ قرمز نسبت به الگوریتم Q و الگوریتم سارسا دارد. علت این امر دید رو به جلو و عقب در مسیرهای شایستگی است در صورتی که الگوریتم Q و سارسا تنها دید رو به عقب دارند.



شکل ١٤. مقایسه سه الگوریتم در میزان تعداد خودروها پشت چراغ قرمز کل چهارراه

جدول ١. مقایسه سه الگوریتم یادگیری تقویتی در سیستم کنترل ترافیکی یک تقاطع

مقایسه سه الگوریتم یادگیری تقویتی	الگوریتم یادگیری مسیرهای شایستگی	الگوریتم یادگیری Q	الگوریتم یادگیری سارسا
میانگین تعداد خودرو پشت چراغ قرمز در هر گام(دستگاه)	٨/١	١٢/٦	١٥/٢
میانگین تاخیر پشت چراغ قرمز در هر گام(ثانیه)	١٢	٢٠/٥	٢٤/٣٧

طبق جدول ١ مشاهده می‌شود که سیستم کنترل ترافیک با الگوریتم مسیرهای شایستگی می‌تواند تعداد خودروهای منتظر و میزان تاخیر زمانی هر خودرو به صورت نسبت به سیستم کنترل ترافیک با الگوریتم یادگیری Q و سیستم کنترل ترافیک با الگوریتم سارسا کاهش دهد.

٥- نتیجه‌گیری

مساله کنترل ترافیک و مینیمم کردن زمان انتظار در تقاطع یکی از دغدغه‌های مهم شهری است. در این مقاله بر اساس سه مد کنترلی بر پایه یادگیری تقویتی بنا شده است. در ابتدا مدل یک چهارراه پویا، متغییر با زمان و تصادفی را شبیه‌سازی کردیم، سپس برای هر تقاطع با استفاده از سه روش الگوریتم یادگیری Q، سارسا و مسیرهای شایستگی، عملیات کنترل چراغ‌ها را

انجام دادیم. ذکر این نکته که در شبیه‌سازی تقاطع، اغتشاش‌ها و جریان‌های ترافیکی مختلفی از جمله تصادف و ساعات اوج ترافیک لحاظ شده تا تطابق بیشتری با مدل واقعی داشته باشد. نتایج نشان می‌دهد که الگوریتم مسیرهای شایستگی به علت داشتن خاصیت دید رو به جلو و دید رو به عقب به صورت همزمان، عملکردی به مراتب بهتری از الگوریتم یادگیری Q و سارسا دارد زیرا الگوریتم یادگیری Q و سارسا فقط خاصیت دید رو به جلو دارند.این مساله باعث می‌شود، تعداد خودروهای پشت چراغ قرمز چهارراه هوشمند با الگوریتم مسیرهای شایستگی نسبت به دو الگوریتم دیگر کمتر باشد.

٦-مراجع

-Chu, T., Qu, S., & Wang, J. (2016). "Large-scale traffic grid signal control with regional Reinforcement Learning". In American Control Conference (ACC), pp. 815-820.

-Dusparic, I., Monteil, J., & Cahill, V. (2016), "Towards autonomic urban traffic control with collaborative multi-policy reinforcement learning". In Intelligent Transportation Systems (ITSC), IEEE 19th International Conference on, pp. 2065-2070.

-Prabuchandran, K. J., AN, H. K., & Bhatnagar, S. (2014), "Multi-agent reinforcement learning for traffic signal control. In Intelligent Transportation Systems (ITSC)", IEEE 17th International Conference on, pp. 2529-2534.

-Wu, L., Zhang, X., & Shi, Z. (2010). "An intelligent fuzzy control for crossroads traffic light". In Intelligent systems (GCIS), second WRI global congress on, Vol. 3, pp. 28-32.

-Odeh, S. M. (2015), "Hybrid algorithm: fuzzy logic-genetic algorithm on traffic light intelligent system". In Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, pp.1-7.

-Abhishek, K., & Misra, B. B. (2016), "Hybrid Genetic Algorithm and time delay neural network model for Forecasting Traffic flow". In Engineering and Technology (ICETECH), IEEE International Conference on, pp. 178-183.

-Chang-sheng, Z., Jian-bo, L., Wen-yi, F., & Xu-gang, M. (2010), "Intelligent dispatch for public traffic vehicles based on improved Genetic –Algorithm". In Future Computer and Communication (ICFCC), 2nd International Conference on, Vol. 3, pp. V3-372.

-Kim, S. (1994), "Application of Petri Networks and Fuzzy Logic to Advanced Traffic Management Systems". Ph.D Thesis, Polytechnic University, USA, 139 P.

-Qiao, J., Yang, N., & Gao, J. (2011), "Two-stage fuzzy logic controller for signalized intersection". IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 41(1), pp.178-184.

-Liu, Z. (2007), "A survey of intelligence methods in urban traffic signal control". IJCSNS International Journal of Computer Science and Network Security, 7(7), pp.105-112.

-Park, B., Messer, C. and Urbanik II, T. (2000), "Enhanced genetic algorithm for signal-timing optimization of oversaturated intersections" Transportation Research Record: Journal of the Transportation Research Board, (1727), pp.32-41.

-Xu Dongling, et al. (1992), "A Fuzzy Controller of Traffic Systems and Its Neural Network Implementation". Information and Control, 21(2): pp. 74-78.

-Tian, Y., Li, Z., Zhou, D., Song, J., & Xiao, D. (2008), "Interactive signal control for over-saturated arterial intersections using fuzzy logic". In Intelligent Transportation Systems,. ITSC. 11th International IEEE Conference on, pp. 1067-1072.

-Sutton RS, Barto AG.(1998), "Reinforcement learning: An introduction". Cambridge: MIT press.